DOCUMENT RESUME

ED 370 351 FL 021 919

AUTHOR de Jong, John H. A. L.; Stoyanova, Fellyanka
TITLE Theory Building: Sample Size and Data-Model Fit.
PUB DATE Mar 94
NOTE 16p.; Paper presented at the Annual Language Testing Research Colloquium (March 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Dutch; Foreign Countries; *Item Response Theory; *Language Research; Language Tests; Listening Comprehension; Models; *Research Methodology; *Sampling; Second Languages; *Statistical Analysis; *Testing; Uncommonly Taught Languages

ABSTRACT

A study of item response theory in language testing research investigated the influence of sample size on (1) the statistical test of data-model fit and (2) the invariance of parameter estimates. Data were drawn from a 1993 administration of the examination of Dutch as a second language to about a thousand candidates, using results from only the listening comprehension segment. One group of examinees was divided into several randomly assembled subsamples, differing in size. Independent analyses of the subsamples were run to assess sample size influence on output variables. Second, a subset of test items was selected for which statistical model fit could be shown, and stability of data-model fit and invariance of item and person parameters over several randomly drawn subsamples was evaluated. Results indicate that estimates of data-model fit and item and person parameters are highly dependent on sample size, but that estimates will be stable for randomly drawn subsamples from a large sample for which statistical fit can be shown. Implications for theory building in language testing research are discussed. Contains 14 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

THEORY BUILDING: SAMPLE SIZE AND DATA-MODEL FIT

ED 370 351

John H.A.L. de Jong
CITO, Arnhem

and

Fellyanka Stoyanova
University of Sofia

Item Response Theory (IRT) is often used in language testing and language testing research. The properties of IRT models such as the Rasch model offer many advantages in the construction of measurement instruments and in the building of theoretical models or constructs of language proficiency. An important requirement is that for these properties to apply, the degree of fit of a data set to the model must be assessed. In this approach, as in any statistical testing of a hypothesis, the objective of the researcher is to falsify the model. The researcher sets up a hypothesis, which (s)he believes to be true. Subsequently (s)he collects data in an attempt to prove the model wrong. If this attempt is not successful, i.e., the data do fit the model, then the researcher can safely assume that, at least for the data used, there is no need to reject the model.

This research was undertaken to investigate the influence of sample size on (1) the statistical test of data-model fit and, (2) the invariance of parameter estimates. Data from the July 1993 administration of the Examination of Dutch as a second language to about one thousand candidates were collected. The examinations comprise separate tests for reading, listening, writing and speaking. For the present study the results of Listening Comprehension test were used. One group of examinees was divided in several randomly assembled subsamples, differing in size. Independent analyses of the subsamples are run to assess the influence of sample size on the output variables. Secondly, a subset of items from the total test was selected for which statistical model fit could be shown and stability of data-model fit and invariance of item and person parameters over several randomly drawn subsamples was evaluated.

The results of this study show that estimates of model-data fit and item and person parameters is highly dependent on sample size, but that estimates will be stable for randomly drawn subsamples from a large sample for which statistical fit can be shown.

The advantages of IRT are a direct result of the strong assumptions underlying IRT models. However, in language testing research the sample sizes used by researchers are often so small that it is highly unlikely that the measurement model can be falsified. A serious implication of the present study then would be, that many reported studies in language testing research should be reconsidered as to their claims with respect to theory building.

Item Response Theory (IRT) is often used in language testing and language testing research. This is because the properties of IRT models offer many advantages for constructing measurement instruments and building theoretical models or constructs of language proficiency. Successful application of IRT, however, is possible only when all assumptions underlying a particular IRT model are met. In other words, the theoretical advantages apply only if the chosen model actually fits the data.

This is of course a well known fact in theory, but in many practical applications of IRT, reports on model-data fit are missing and the results of a calibration are used without any information on the fit of the chosen model to the data set. In the best of cases the model-data fit study begins, and, ends with a simple report of the test statistic, based on some statistical test of model fit.

FL021919

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John H.A.L. de Jong

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

The model-data fit study, however, requires more than a statistical test of fit, because all known statistical tests are very sensitive to sample size. As Hambleton and Swaminathan (1985, pp. 154-155) have shown, the number of misfitting items increases with increasing sample size. Consequently, if a researcher would wish to insure statistical fit, (s)he would be wise to base the analysis on a small sample. However, when using a small sample, there is only one type of result of statistical modelling which allows clear and categorical conclusions, that is, when the test statistic shows statistical misfit. The conclusion in that case must be that there is little probability for the researcher to make a Type I error, i.e., to reject the model on the basis of the data whereas in actual fact the model does apply. If, on the other hand, the model cannot be rejected on the basis of a small sample data set, the researcher should suspend judgement until more data have been gathered, because the probability of a Type II error, i.e., erroneously accepting the model, depends directly on the power of the statistical test and therefore on the sample size (Hambleton, 1989, p.173). In other words, statistical model fitting is justifiable only for the purpose of rejecting the hypothesis of model-data fit, but is an insufficient basis to decide on acceptance of the hypothesis of model-data fit.

Sample size, however, does not only influence the statistical model test, but all other results of the calibration too. For example, Hambleton and Cook (1983, pp. 43-46) have shown that sample size has a substantial impact on the precision of the size of the standard error of estimates. From this point of view they recommended sample sizes of over 200 examinees. In a simulation study Ree and Jensen (1983) analyzed the effect of sample size on the intercorrelation of known and estimated item parameters using a three parameter model and confirmed that stable estimates of item discrimination and guessing parameters can be attained only with sample sizes of over 2000. Tang, Way and Carey (1993) report similarly poor stability of item parameters, especially the a- and c-parameters, for sample sizes of less than 2000.

Apparently unaware of these basic requirements from statistical theory and their confirmation by empirical findings, researchers in many practically oriented studies in the field of language testing use extremely small samples to prove their structural theories. To our knowledge, none of these studies is supported by practically oriented studies analyzing the invariance of parameter estimates based on small samples.

In this paper we present the results of a study, which analyzes the effect of the sample size on a) the statistical test of data-model fit and, b) on the invariance of parameter estimates.

We have based our study on the Rasch model. Among the family of IRT models the one parameter Rasch Model has been shown to yield stable estimates with smaller samples than are needed using other IRT models (Boldt, 1994; Lord, 1983). Therefore, if the effect of sample size on statistical tests of fit and on the invariance of parameters can be shown to be substantial for the Rasch model, than it will certainly effect these statistics in other IRT models.

METHOD

Subjects

The total sample was formed by 973 candidates taking the test of Listening Comprehension Test in the July 1993 test form of Examination I of Dutch as a second language test. For the first part of the study all subjects in this total sample (ST) were randomly divided over five subsamples (S1,..., S5) increasing in size. Furthermore, all subjects in subsamples 1 and 2 (S1 and S2) were joined to form a new subsample (S12) making it equal in size to subsample 4, and all subjects in subsample 4 were randomly redistributed over two new subsamples (S4a and S4b) in such a way to make them equal in size to subsamples 1 and 2 respectively. For the second part of the study 43 subjects with incomplete records (skipped items) were removed from the data file. The remaining 930 candidates formed a selection of the total sample (STs). Subsequently these were randomly divided over two subsamples of equal size. A complete overview of all (sub)samples is provided in Table 1.

Table 1
Overview of Samples: Names, Codes and Sizes

Sample Name	Sample Code	Sample Size
Total Sample	ST	973
Subsample 1	S1	100
Subsample 2	S2	150
Subsample 3	S3	200
Subsample 4	S4	250
Subsample 5	S5	273
Subsample 1 + 2	S12	250
Subsample 4A	S4a	100
Subsample 4B	S4b	150
Selected Total	STs	930
First Half Selected Total	S½s.1	465
Second Half Selected Total	S½s.2	465

Instruments

All examinees took the Listening Comprehension test of the July 1993 version of the Test of Dutch as a Second Language, prepared by the Language Department of CITO. The July 1993 form of the test consists of 46 multiple-choice items, scored dichotomously and aims at assessing the ability of examinees to understand Dutch spoken spontaneously in settings such as interviews, instructions, news, and messages.

Procedures

Data of all samples were analyzed according to the One Parameter Logistic Model implemented in the computer program OPLM (Verhelst et al., 1993). OPLM is a member of the class of one parameter models proposed by Rasch (1960) but allows the researcher to plug in integer discrimination parameters. While thus providing a means to weigh items differentially, full advantage is taken of the property of sufficient statistics available in Rasch models for the estimation of item difficulty parameters and thus allowing for statistically sound tests of data-model fit.

For each subsample dataset, the Rasch model item parameters were estimated by a conditional maximum likelihood (CML) method. Given the estimate of the item parameters, the ability parameters were estimated using an unrestricted maximum likelihood method. Model fit of the data was evaluated using a method proposed by Glas (1988; 1989). The method is based on a suggestion for a test of fit by Martin Löf (1973), but does not make the assumption that the number of subjects obtaining a particular score is a Poisson-distributed random variable. The method provides an overall test of fit and also a means for identifying persons and items contributing to lack of model fit. Model fit can be tested using score groups, i.e., subsets of persons obtaining equal scores. For long tests, with a large number of score groups, adjacent score groups can be combined in subgroups, each containing an approximately equal number of subjects from the total sample.

The principle on which the test of fit is based can be summarized as follows. Suppose the test consists of k items, and G subgroups are formed. For every subgroup (indexed g, so $g = 1,\ldots, G$) and every item (indexed i, so $i = 1,\ldots, k$) the number of correct responses is computed, these frequencies are denoted by m_{gi}. Let $E_c(m_{gi})$ stand for the conditional expectation of m_{gi}, evaluated using conditional maximum likelihood of the item parameters. Then the deviances, $d_{gi} = m_{gi} - E_c(m_{gi})$ can be used to determine item fit. To evaluate the size of the deviances, a scaled deviance, $d^*_{gi} = d_{gi}/(var_c(m_{gi}))^{½}$, is computed, which has an approximately standard normal distribution.

To obtain a formal measure of overall model fit, all deviances must be combined into a quadratic form, indexed R1c, which has an asymptotic chi-square distribution. For a detailed description of this global fit statistic one is referred to Martin Löf (1973) or Glas (1988; 1989). For every item the scaled deviances can be squared and summed over subgroups to obtain an index of item fit which has an approximate chi-square distribution with a number of degrees of freedom equal to the number of score groups minus one.

For the first part of the study test calibration runs were made separately for the total sample and for all subsamples (excluding S1/2s subsamples). In this way 9 different sets of item and ability estimates were obtained. After converting the separate estimates to a common scale, the results of these calibrations were checked on two aspects:

1) the stability of the model-data fit statistics for the different calibrations;
2) the invariance of item and ability estimates.

For the second part of the study items were deleted in an iterative procedure described by De Jong (1983) and De Jong and Glas (1987) in order to obtain a subset of items which formed a Rasch scale. Subsequently test calibration runs were made for the total sample, for the selected sample and both halves of the selected sample. Finally, calibration runs were made using all samples used in the first part of the study.

RESULTS AND DISCUSSION:

1. THE EFFECT OF THE SAMPLE SIZE ON THE STATISTICAL TEST

Table 2 presents classical test statistics analysis, the statistics and the level of significance of the statistical test for model-data fit and the number of misfitting items for different subsamples.

Table 2:
Results of calibration runs on total test for different sample sizes

Sample	N	Mean (raw score)	SD	Alpha[a]	R1c[b]	p	Number of misfitting items($p < .05$)
S1	100	30.02	7.03	.82	88.675	.5197	2
S2	150	29.42	7.87	.86	137.348	.0339	6
S3	200	31.04	7.95	.87	232.577	.0000	7
S4	250	30.52	7.89	.87	247.780	.0000	7
S5	273	30.47	7.76	.86	295.246	.0000	10
ST	973	30.39	7.80	.86	568.172	.0000	22
S4A	100	31.96	7.44	.86	137.348	.0010	8
S4B	150	29.55	8.04	.87	202.461	.0001	6
S12	250	29.66	7.55	.85	188.315	.0013	6

a: Cronbach's Alpha
b: R1c is a statistic of total model-data fit

The results presented in Table 2 confirm the results reported by Hambleton and Swaminathan (1985, pp. 154-155) and demonstrate the effect of the sample size on the number of misfitting items. Likewise, with increasing sample size the probability of total model-data misfit increases too. As can be seen, the data do not fit the model even for S12 which is simple union of S1 and S2, where for each of these statistical model-data fit is apparent.

The comparison of the statistics for total model-data fit shows that there are two pairs of samples with equal sizes (S1-S4A and S2-S4B) in each of which one

sample leads to a different conclusion than the other if the outcome of the statistical test of model-data fit is taken as a sufficient proof of actual fit. If a study were based on either sample S1 or S2 the conclusion of the naive researcher would be that the model fits the data. If, on the other hand either sample S4A or S4B were used this same researcher would have to conclude to the opposite. In this way, depending on the desired outcome anything can be proved. This result shows that the problem of model-data fit goes beyond the statistical test of fit.

The number of misfitting items in the different analyses also shows a substantial variance. Only 16 out of the total 46 items fit for all 9 analyzed samples and only item 1 misfits in all analyses. The remaining 29 items, 64% of the total number, fit in some analyses and show misfit in others. Ultimately this means that the persistent researcher could always find a sample providing statistical support for rejecting or accepting any item.

Table 3 presents the results of the first set of calibration runs for the second part of the study. First the results of total sample (ST) based on the complete set of items is reproduced from Table 2 for easy comparison. Next the results of a calibration after removing subjects with incomplete records is presented (STs). Removal of incomplete records did not result in any substantial changes of the statistics reported. The following three rows in Table 3 present the results of calibrations of a selected subset of items forming a Rasch scale, first for the total group (minus incomplete records) and secondly, for calibrations based on data from the randomly assembled two halves of STs (STs½.1 and STs½.2).

Table 3
Results of calibration runs on total test and selected subtest for different sample sizes

Sample	N	Mean (raw score)	SD	Alpha[a]	R1c[b]	p	Number of misfitting items($p < .05$)
Total test (k = 46)							
ST	973	30.39	7.80	.86	568.172	.0000	22
STs	930	30.56	7.82	.86	548.708	.0000	21
Selected subtest (k = 21)							
STs	930	14.59	4.33	.82	79.430	.0473	1
S½s.1	465	14.76	4.32	.82	77.248	.0663	2
S½s.2	465	14.41	4.32	.81	79.146	.0495	0

a: Cronbach's Alpha
b: R1c is a statistic of total model-data fit

The results presented in Table 3 show that once a Rasch scale has been formed basic statistics are independent of sample size or composition, mean scores,

standard deviations, reliability indices (alpha), item fit statistics, and number of misfitting items remain essentially invariant over calibrations.

To further substantiate these findings the results of calibration runs using the subset of items that form a Rasch scale on all samples used in the first part of the study are presented in Table 4.

Table 4
Results of calibration runs on selected subtests for different sample sizes

Sample	N	Mean (raw score)	SD	Alpha[a]	R1c[b]	p	Number of misfitting items($p < .05$)
S1	100	14.65	3.93	.77	46.27	.2294	1
S2	150	13.85	4.21	.80	41.15	.4200	1
S3	200	14.99	4.35	.83	78.93	.0512	3
S4	250	14.58	4.40	.82	42.67	.9558	2
S5	273	14.35	4.43	.82	71.91	.1396	1
ST	973	14.49	4.34	.82	84.72	.0195	3
S4A	100	15.31	4.17	.82	15.96	.7193	1
S4B	150	14.09	4.48	.82	47.56	.8776	1
S12	250	14.17	4.12	.79	60.10	.4719	1

a: Cronbach's Alpha
b: R1c is a statistic of total model-data fit

Table 4 clearly shows that all basic statistics remain essentially invariant once the Rasch model has been shown to fit the data.

2. INVARIANCE OF ITEM AND ABILITY PARAMETERS

The application of Item Response Modelling makes sense only in the case of model-data fit. That is why in the next analysis we will use the results of the calibrations based only on the first two subsamples, for which there is at least a statistical proof of model-data fit. If the reported fit-indexes were indeed sufficient ground for acceptance of the hypothesis of model-data fit, the model's properties should hold and item and ability parameters should be invariant.

A method to checking model features is to divide a sample randomly into two subsamples and compare the results of the separate calibrations of these subsamples. Figure 1a presents the scatterplot of estimated b-values, based on the calibrations of two random subsamples from Sample 1, and Figure 1b presents these results for subsamples from Sample 2. Clearly the b-value estimates are far

from being invariant. Note ,however, that for the subsamples from Sample 2, which contain more subjects, the correlation between b-values is larger (r_{S1} = 0.749 and r_{S2} = 0.897).

Figures 1a, 1b

The same procedure applied to a sufficiently large sample gives different results. Figure 2 presents a scatterplot of item parameter estimates based on calibrations of the two randomly assembled two halves of STs, STs½.1 and STs½.2. The figure demonstrates that the feature of invariance of item parameters holds when statistical model-data fit is achieved for a sufficiently large sample.

Figure 2

A direct comparison of the estimates of b-values based on the calibrations in Samples 1 and 2 also shows a fair amount of variation. The differences between the two sets of b-value estimates for all test items are plotted on Figure 3. The differences vary between -0.85 and +1.27 which can hardly be taken as a proof of feature of invariance.

Figure 3

To illustrate the impact of the differences as presented in Figure 3, suppose a subject taking the test solves item 1 correctly and item 12 incorrectly. On the basis of IRT we would conclude that the ability of this subject is somewhere between b_1 and b_{12}. If we used item parameter estimates based on Sample 1, this would imply that the ability of this subject is in the interval [-1.35;-1.29], but using estimates based on Sample 2 it would mean an ability estimate in the interval [-2.62; -1.73]. Figure 4 presents the item characteristic curves of items 1 and 12 based on the calibration of both Sample 1 and Sample 2. The difference between the two ability intervals clearly shows that the feature of invariance of ability estimates doesn't apply.

Figure 4

Another attractive feature of IRT is that the estimate of a subject's ability would be the same if different subsets of items from a test were used. If the significant statistical level of fit for the calibrations of Sample 1 and Sample 2 were sufficient proof of true model-data fit, then it would hold for ability estimates based on these subsets of items. However, the examples seem to indicate that the feature does not hold and the next two studies provide further substantiation.

Firstly, we divided the test into two parts, where Part 1 comprises the first half of the test, and Part 2 the second half. For each part a separate set of ability parameters was estimated based on the calibrations of Sample 1 for Part 1 and on Sample 2 for Part 2. The ability of all examinees in a third sample (Sample 3) was estimated using both sets of ability parameter estimates. Figure 5 provides a scatterplot of the abilities estimates and shows that the ability estimates are far from being invariant. The correlation of the two estimates is 0.737.

Figure 5

Secondly, we divided the test into two parts according to the difficulty estimates of the calibrations of Sample 1. Again the ability of the examinees of Sample 3 was estimated twice, using the Easy and Hard Subtest respectively. Figure 6 presents the scatterplot of ability estimates of subjects in Sample 3 based on Easy and Hard Subtest calibrations of subjects in Sample 1. As could be seen the ability estimates vary substantially over both calibrations. The correlation coefficient is respectively 0.750, indicating that the basic advantage of the item response modelling does not apply in this case.

Figure 6

In other words, in spite of the statistical fit the model does not truly fit the data. The main reason for this discrepancy is the small size of the samples, which cannot insure stable estimates of item and ability parameters.

CONCLUSIONS:

The application of IRT provides many advantages to researchers in the field of testing. Before application, however, the researcher has to be sure that the data fit the chosen model. A statistical test of fit of the model as such is insufficient to accept the hypothesis of model-data fit. The results of our study shown, that:

1. For one and the same test it is always possible to find samples with size < 200 for which the statistics will allow to accept both hypotheses: H_0

and H_1;

2. Statistical fit for small samples does not insure invariance of item and ability parameter estimates;
3. Statistical fit for large samples allows generalization to samples varying in size and confidence that model features apply.

In language testing research the sample size used is often too small. Moreover, model-data fit studies are usually missing, and in the best of cases begin and end by simply reporting a fit statistic. However, the results of studies based on Item Response Modelling are used for decision making and general conclusions. The results of this study show that many studies in language testing research need reconsidering as to their claims with respect to theory building.

REFERENCES:

BOLDT, R.F. (1994) Simulated equating using several item response curves. TOEFL Technical Report (TR-8). Princeton: Educational Testing Service.

DE JONG, J.H.A.L. (1983) Focusing in on a latent trait : An attempt at construct validation by means of the Rasch model. In: J. van Weeren (ed.) *Practice and Problems in Language Testing 5*. Arnhem: Cito.

DE JONG, J.H.A.L. & C.A.W. Glas (1987) Validation of listening comprehension tests using item response theory. *Language Testing, 4*, 170-194.

GLAS, C.A.W. (1988) The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

GLAS, C.A.W. (1989) *Contributions to estimating and testing Rasch models.* PhD Dissertation. Enschede: Twente University.

HAMBLETON,R.K. (1989) Principles and Selected Applications of Item Response Theory. In: R. Linn (ed.) *Educational Measurement.* Third Edition. New York: American Council on Education and Macmillan Publishing Company.

HAMBLETON, R.K. L. COOK. (1983). Robustness of Item Response Models and Effects of Test Length and Sample Size on the Precision of Ability Estimates. In: *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.

HAMBLETON, R.K. & H. SWAMINATHAN (1985). Item Response Theory: Principles and Applications. Kluwer. Nijhoff Publishing.

LORD, F.M. (1983) Small N justifies Rasch model. In: D.J. Weiss (ed.) *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.

MARTIN LÖF, P. (1973: *Statistika modeller. Anteckningar från seminarier Lasåret 1969-1970*; Utarbetade av Rolf Sunberg obetydligt ändrat nytryk, oktober 1973. Stockholm: Institutet för Försäkringsmatamatik och Matamatisk Statistik vid Stockholms Universitet.

RASCH, G. (1960) *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.

REE, M.J. AND H.E. JENSEN (1983) Effect of Sample Size on Linear Equating of Item Characteristic Curve Parameters. In: D.J. Weiss (ed.) *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.In: New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. Ed. by D. Weiss. Academic Press, 136-146.

Tang, K.L., W.D. Way & P.A. Carey (1993) The effect of small calibration sample sizes on TOEFL IRT-based equating. TOEFL Technical Report (TR-7). Princeton: Educational Testing Service.

VERHELST, N.D. ET ALL. (1993) *OPLM: One Parameter Logistic Model. Computer Program and Manual.* (Preliminary Version) CITO, Arnhem.

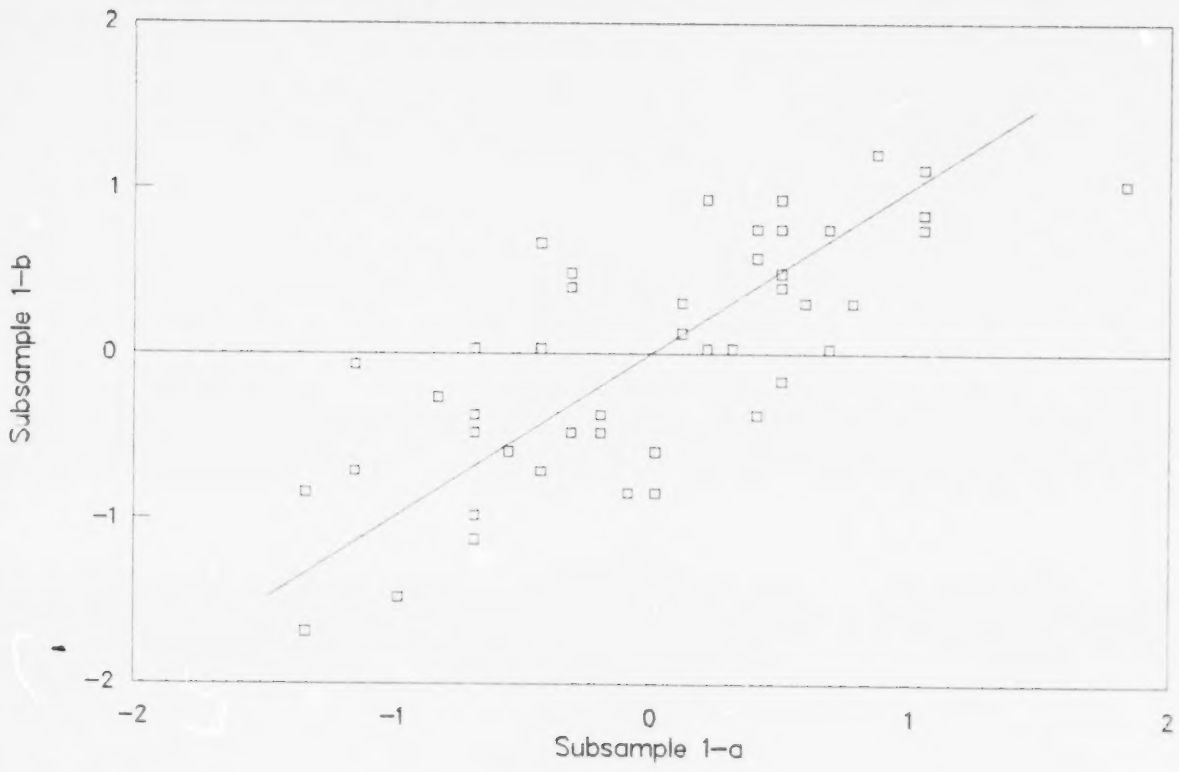


Figure 1a

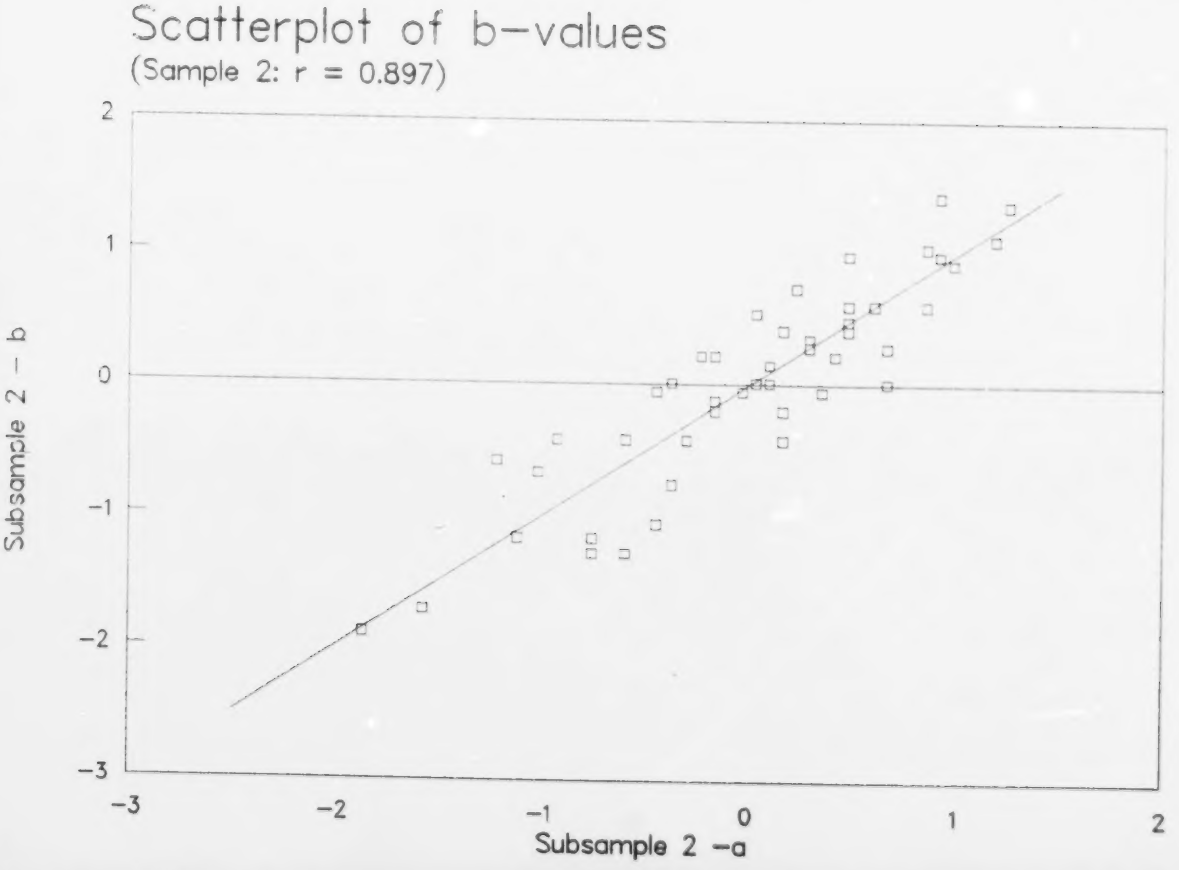


Figure 1b

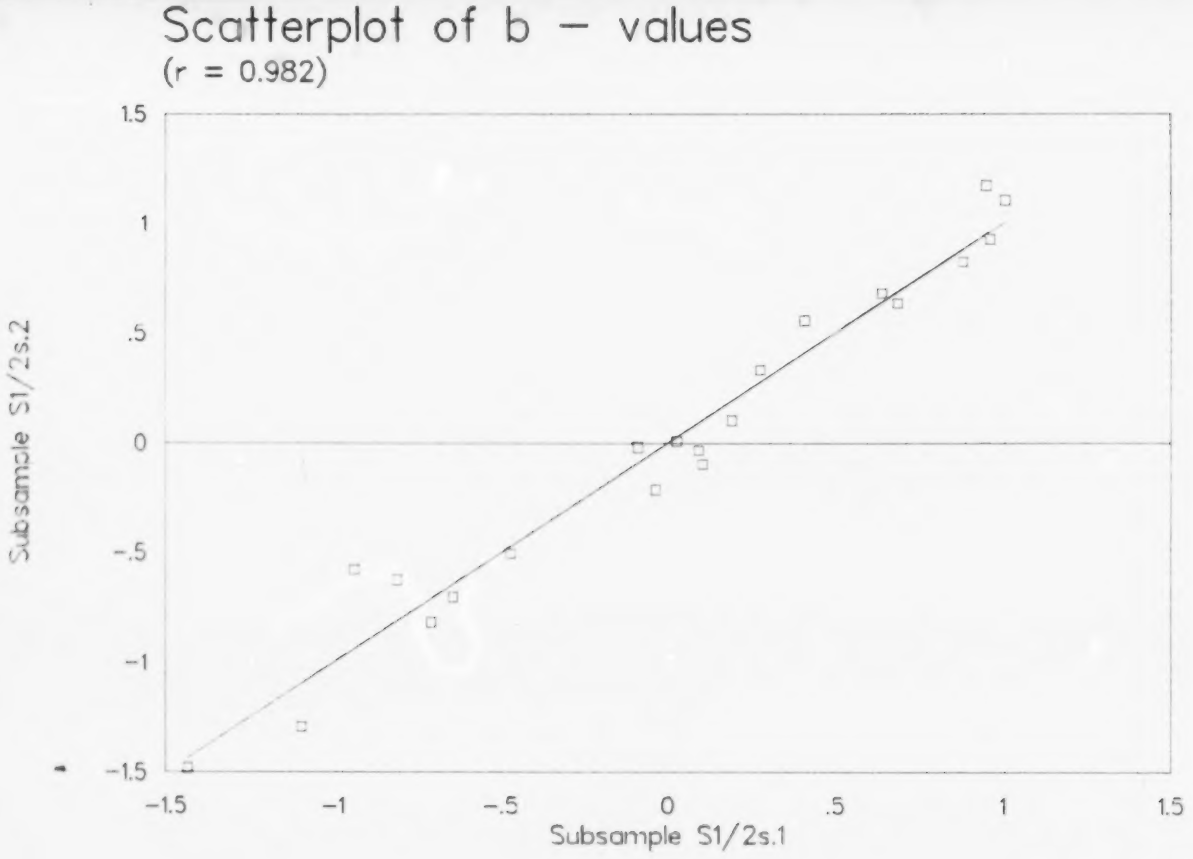


Figure 2

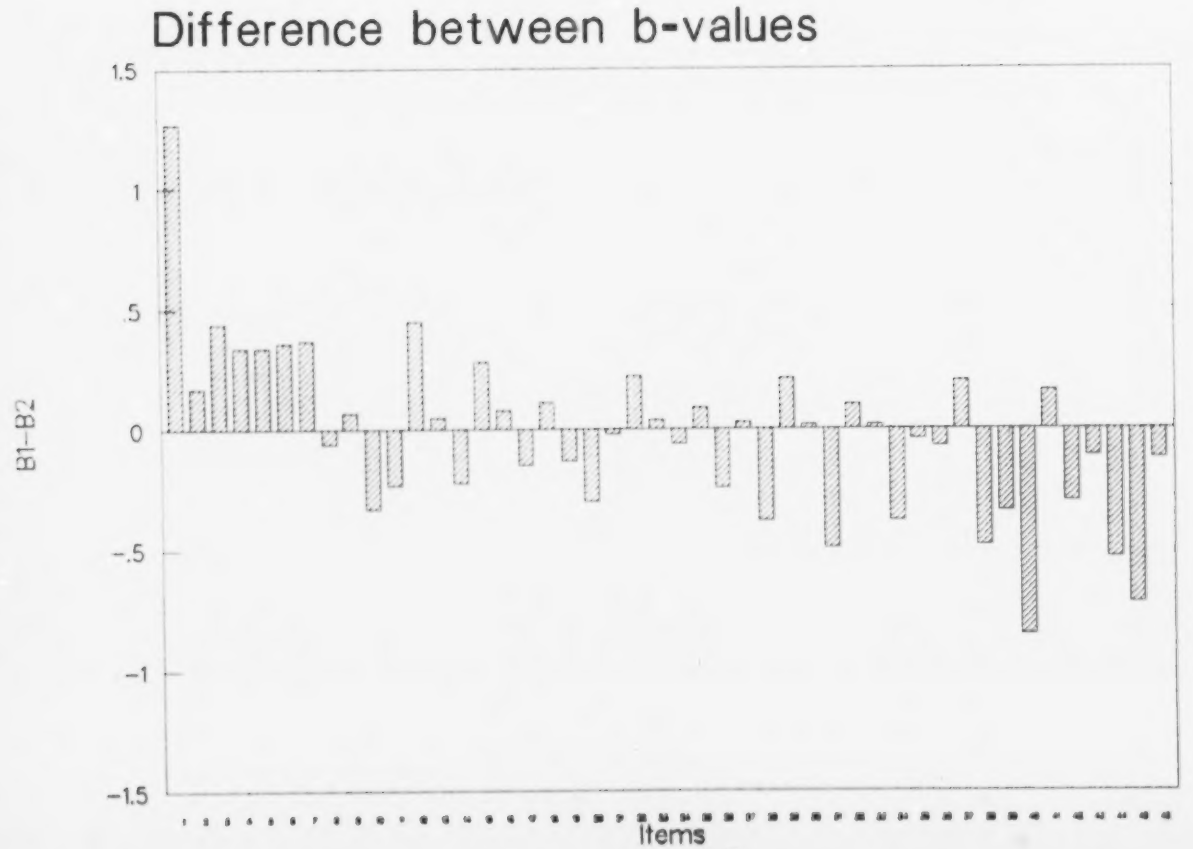


Figure 3

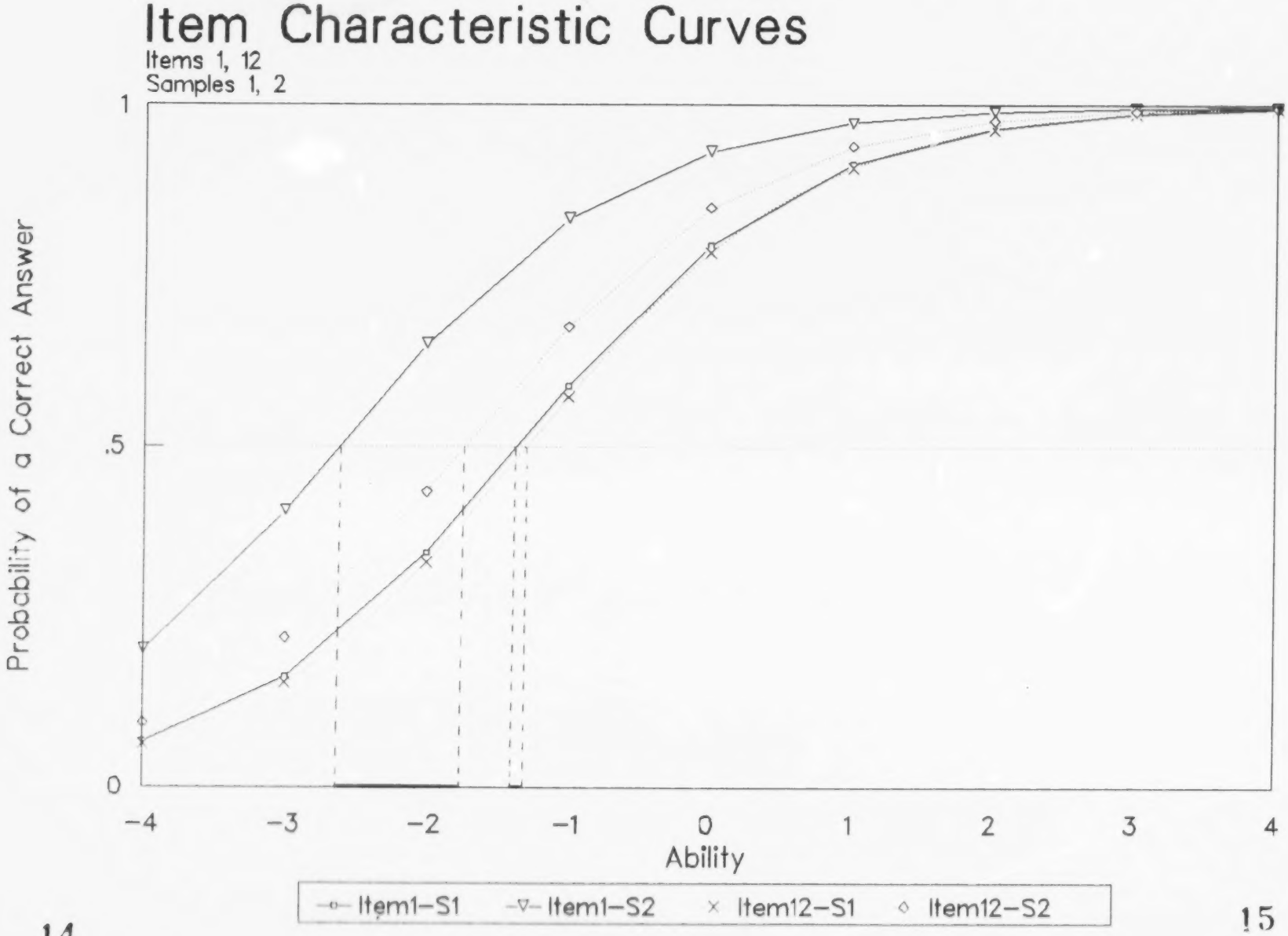

Item Characteristic Curves
Items 1, 12
Samples 1, 2
Probability of a Correct Answer
1
.5
0
−4
−3
−2
−1
0
1
2
3
4
Ability
Item1−S1
Item1−S2
Item12−S1
Item12−S2

Figure 4

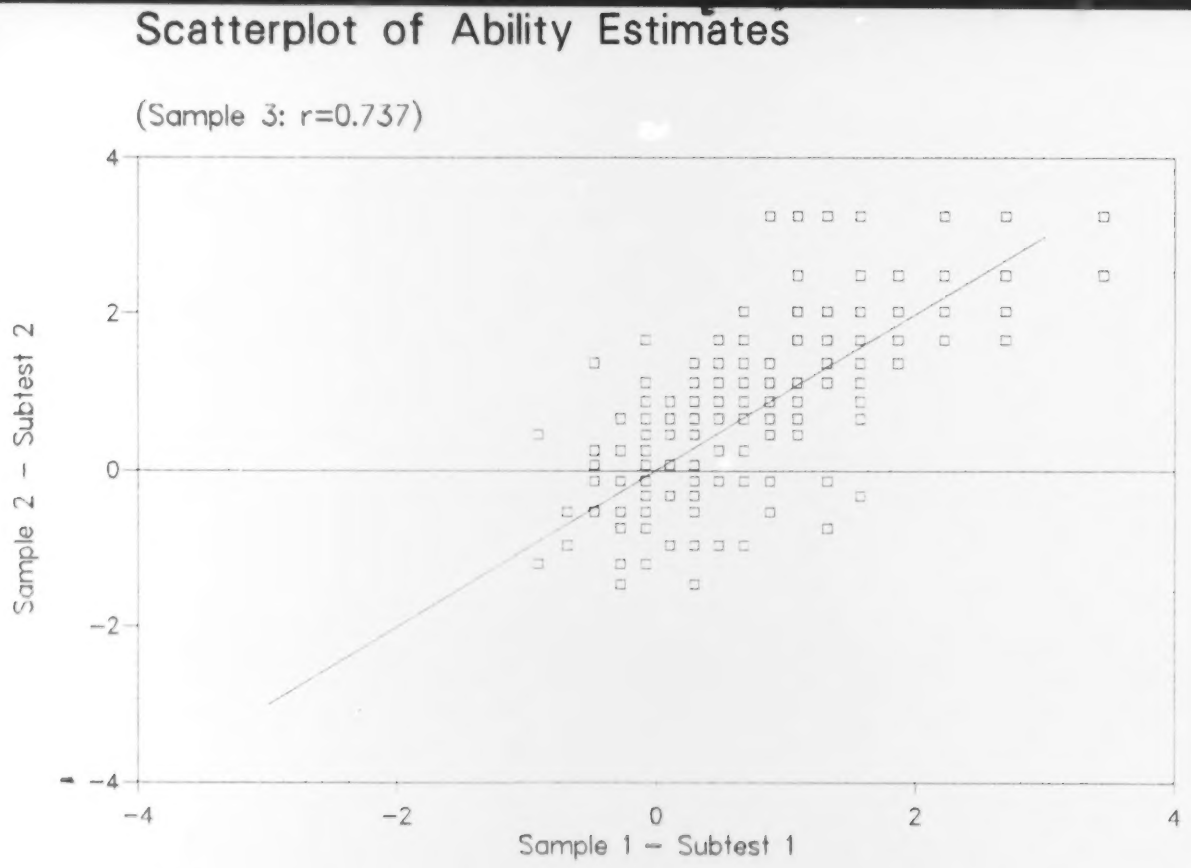
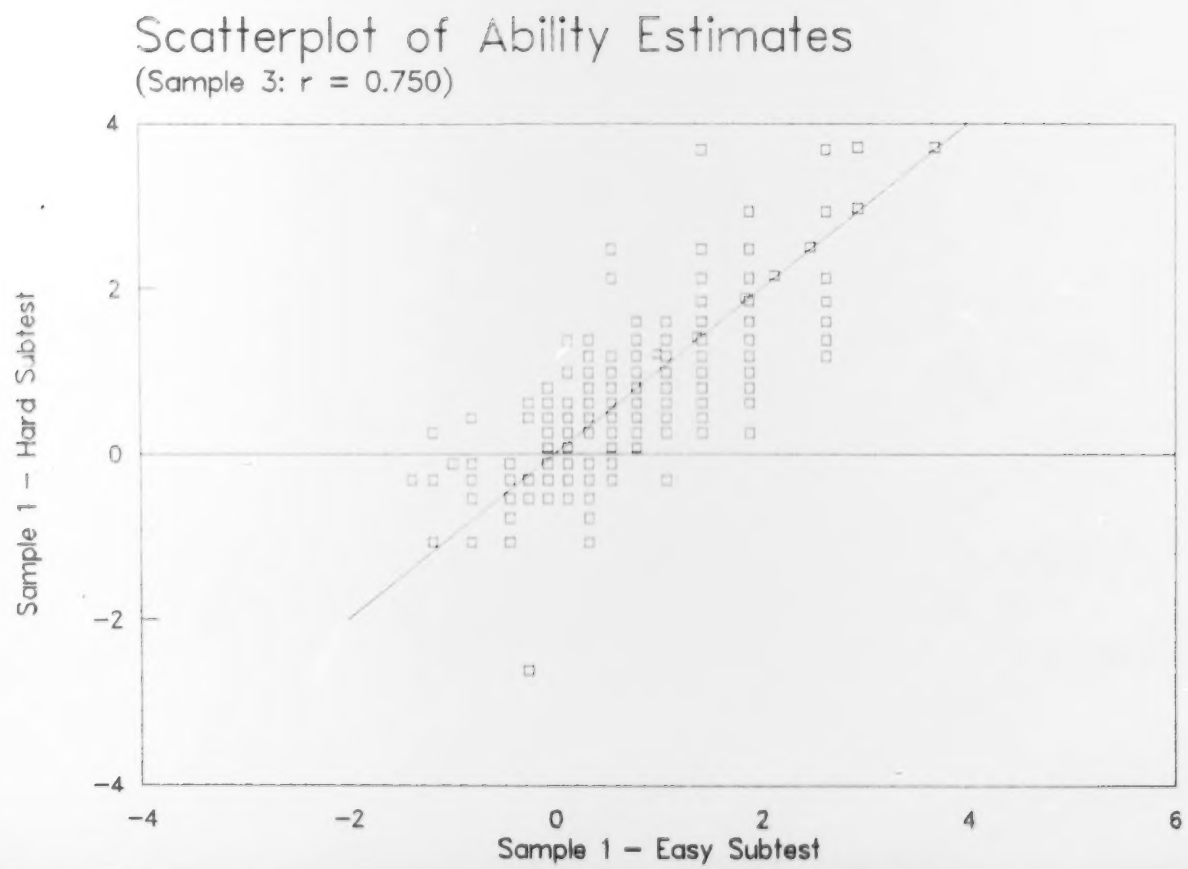

Scatterplot of Ability Estimates
(Sample 3: r=0.737)
Sample 2 – Subtest 2
Sample 1 – Subtest 1
4
2
0
−2
−4
−4
−2
0
2
4

Figure 5


Scatterplot of Ability Estimates
(Sample 3: r = 0.750)
Sample 1 – Hard Subtest
Sample 1 – Easy Subtest
4
2
0
−2
−4
−4
−2
0
2
4
6